Long Memory of Financial Time Series and Hidden Markov Models with Time-Varying Parameters

Nystrup, Peter; Madsen, Henrik; Lindström, Erik

Published in:
Journal of Forecasting

Link to article, DOI:
10.1002/for.2447

Publication date:
2016

Document Version
Peer reviewed version

Long memory of financial time series and hidden Markov models with time-varying parameters

Peter Nystrup[ab]*, Henrik Madsen[b], and Erik Lindström[c]

[a] *Sampension, Denmark*
[b] *Department of Applied Mathematics and Computer Science, Technical University of Denmark*
[c] *Centre for Mathematical Sciences, Lund University, Sweden*

Abstract

Hidden Markov models are often used to model daily returns and to infer the hidden state of financial markets. Previous studies have found that the estimated models change over time, but the implications of the time-varying behavior have not been thoroughly examined. This paper presents an adaptive estimation approach that allows for the parameters of the estimated models to be time varying. It is shown that a two-state Gaussian hidden Markov model with time-varying parameters is able to reproduce the long memory of squared daily returns that was previously believed to be the most difficult fact to reproduce with a hidden Markov model. Capturing the time-varying behavior of the parameters also leads to improved one-step density forecasts. Finally, it is shown that the forecasting performance of the estimated models can be further improved using local smoothing to forecast the parameter variations.

Keywords: Hidden Markov models; Daily returns; Long memory; Adaptive estimation; Time-varying parameters

Introduction

Many different stylized facts have been established for financial returns, see e.g. Granger and Ding (1995a,b), Granger et al. (2000), Cont (2001), and Malmsten and Teräsvirta (2010). Rydén et al. (1998) showed the ability of a hidden Markov model (HMM) to reproduce most of the stylized facts of daily return series introduced by Granger and Ding (1995a,b). In an HMM, the distribution that generates an observation depends on the state of an unobserved Markov chain. Rydén et al. (1998) found that the one stylized fact that could not be reproduced by an HMM was the slow decay of the autocorrelation function (ACF) of squared and absolute daily returns, which is of great importance in financial risk management. The daily returns do not have the long-memory property themselves, only their squared and absolute values do. Rydén et al. (1998) considered this stylized fact to be the most difficult to reproduce with an HMM.

According to Bulla and Bulla (2006), the lack of flexibility of an HMM to model this temporal higher-order dependence can be explained by the implicit assumption of geometrically distributed sojourn times in the hidden states. This led them to consider hidden semi-Markov models (HSMMs) in which the sojourn time distribution is modeled explicitly for each hidden state so that the Markov property is transferred to the embedded first-order Markov chain. They found that an HSMM with

*Correspondence to: Peter Nystrup, DTU Compute, Asmussens Allé, Building 303B, 2800 Kgs. Lyngby, Denmark
Email: pnys@dtu.dk

negative-binomially distributed sojourn times was better than the HMM at reproducing the long-memory property of squared daily returns.

Bulla (2011) later showed that HMMs with t-distributed components reproduce most of the stylized facts as well or better than the Gaussian HMM at the same time as increasing the persistence of the visited states and the robustness to outliers. Bulla (2011) also found that models with three states provide a better fit than models with two states. In Nystrup et al. (2015b), an extension to continuous time was presented and it was shown that a continuous-time Gaussian HMM with four states provides a better fit than discrete-time models with three states with a similar number of parameters.

The data analyzed in this paper is daily returns of the S&P 500 stock index from 1928 to 2014. It is the same time series that was studied in the majority of the above-mentioned studies just extended through the end of 2014. Granger and Ding (1995a) divided the full sample into ten subsamples of 1,700 observations, corresponding to a little less than seven years, as they believed it was likely that with such a long time span there could have been structural shifts in the data-generating process. Using the same approach, Rydén et al. (1998) and Bulla (2011) found that the estimated HMMs, including the number of states and the type of conditional distributions, changed considerably between the subsamples.

HMMs are popular for inferring the hidden state of financial markets and several studies have shown the profitability of dynamic asset allocation strategies based on this class of models (see e.g. Bulla et al., 2011; Nystrup et al., 2015a). The profitability of those strategies is directly related to the persistence of the volatility. It is therefore relevant to explore in depth the importance of the time-varying behavior for the models' ability to reproduce the long memory and forecast future returns. Failure to account for the time-varying behavior of the estimated models is likely part of the reason why regime-switching models often get outperformed by a simple random walk model when used for out-of-sample forecasting, as discussed by Dacco and Satchell (1999).

In this paper, an adaptive estimation approach that allows for the parameters of the estimated models to be changing over time is presented as an alternative to fixed-length forgetting. After all, it is unlikely that the parameters change after exactly 1,700 observations. The time variation is observation driven based on the score function of the predictive likelihood function, which is related to the generalized autoregressive score (GAS) model of Creal et al. (2013).

In agreement with the findings by Rydén et al. (1998) and Bulla (2011), the parameters of the estimated models are found to vary significantly throughout the data period. As a consequence of the time-varying transition probabilities, the sojourn time distribution is not the memoryless geometric distribution. A two-state Gaussian HMM with time-varying parameters is shown to reproduce the long memory of the squared daily returns. Faster adaption to the parameter changes improves both the fit to the ACF of the squared returns and the one-step density forecasts. Using local smoothing to forecast the parameter variations, it is possible to further improve the density forecasts. Finally, the need for a third state or a conditional t-distribution in the high-variance state to capture the full extent of excess kurtosis is discussed in light of the non-stationary behavior of the estimated models.

Section 2 gives an introduction to the HMM. Section 3 discusses the relation between long memory and regime switching. In Section 4, a method for adaptive parameter estimation is outlined. Section 5 contains a description of the data. The results are presented in Section 6 and Section 7 concludes.

The hidden Markov model

In a hidden Markov model, the probability distribution that generates an observation depends on the state of an underlying and unobserved Markov process. An HMM is a particular kind of dependent mixture and is therefore also referred to as a Markov-switching mixture model. General references to the subject include Cappé et al. (2005), Frühwirth-Schnatter (2006), and Zucchini and MacDonald (2009).

A sequence of discrete random variables $\{S_t : t \in \mathbb{N}\}$ is said to be a first-order Markov chain if, for all $t \in \mathbb{N}$, it satisfies the Markov property:

$$\Pr\left(S_{t+1} \middle| S_t, \ldots, S_1\right) = \Pr\left(S_{t+1} \middle| S_t\right). \tag{1}$$

The conditional probabilities $\Pr\left(S_{u+t} = j \middle| S_u = i\right) = \gamma_{ij}\left(t\right)$ are called transition probabilities. The Markov chain is said to be homogeneous if the transition probabilities are independent of u, and inhomogeneous otherwise.

If the Markov chain $\{S_t\}$ has m states, then the bivariate stochastic process $\{(S_t, X_t)\}$ is called an m-state HMM. With $\boldsymbol{S}^{(t)}$ and $\boldsymbol{X}^{(t)}$ representing the values from time 1 to time t, the simplest model of this kind can be summarized by

$$\Pr\left(S_t \middle| \boldsymbol{S}^{(t-1)}\right) = \Pr\left(S_t \middle| S_{t-1}\right), \quad t = 2, 3, \ldots, \tag{2a}$$

$$\Pr\left(X_t \middle| \boldsymbol{X}^{(t-1)}, \boldsymbol{S}^{(t)}\right) = \Pr\left(X_t \middle| S_t\right), \quad t \in \mathbb{N}. \tag{2b}$$

Hence, when the current state S_t is known, the distribution of X_t depends only on S_t. This causes the autocorrelation of $\{X_t\}$ to be strongly dependent on the persistence of $\{S_t\}$.

An HMM is a state-space model with finite state space where (2a) is the state equation and (2b) is the observation equation. A specific observation can usually arise from more than one state as the support of the conditional distributions overlaps. The unobserved state process $\{S_t\}$ is therefore not directly observable through the observation process $\{X_t\}$, but can only be estimated.

As an example, consider the two-state model with Gaussian conditional densities:

$$X_t = \mu_{S_t} + \varepsilon_{S_t}, \quad \varepsilon_{S_t} \sim N\left(0, \sigma_{S_t}^2\right),$$

where

$$\mu_{S_t} = \begin{cases} \mu_1, & \text{if } S_t = 1, \\ \mu_2, & \text{if } S_t = 2, \end{cases} \quad \sigma_{S_t}^2 = \begin{cases} \sigma_1^2, & \text{if } S_t = 1, \\ \sigma_2^2, & \text{if } S_t = 2, \end{cases} \quad \text{and } \boldsymbol{\Gamma} = \begin{bmatrix} \gamma_{11} & 1 - \gamma_{11} \\ 1 - \gamma_{22} & \gamma_{22} \end{bmatrix}.$$

For this model, the value of the autocorrelation function at lag k is

$$\rho_{X_t}\left(k \middle| \theta\right) = \frac{\pi_1\left(1 - \pi_1\right)\left(\mu_1 - \mu_2\right)^2}{\sigma^2}\lambda^k \tag{3}$$

and the autocorrelation function for the squared process is

$$\rho_{X_t^2}\left(k \middle| \theta\right) = \frac{\pi_1\left(1 - \pi_1\right)\left(\mu_1^2 - \mu_2^2 + \sigma_1^2 - \sigma_2^2\right)^2}{\mathrm{E}\left[X_t^4 \middle| \theta\right] - \mathrm{E}\left[X_t^2 \middle| \theta\right]^2}\lambda^k, \tag{4}$$

where π_1 is the stationary probability of state one and $\lambda = \gamma_{11} + \gamma_{22} - 1$ is the second largest

eigenvalue of $\mathbf{\Gamma}$ (Frühwirth-Schnatter, 2006).[1] It is evident from these expressions, as noted by Rydén et al. (1998), that an HMM with constant parameters can only reproduce an exponentially decaying autocorrelation structure. The ACF of the first-order process becomes zero if the mean values are equal whereas persistence in the squared process can be induced either by a difference in the means or by a difference in the variances. In both cases the persistence increases with the combined persistence of the states as measured by λ.

The sojourn times are implicitly assumed to be geometrically distributed:

$$\Pr\left(\text{'staying } t \text{ time steps in state } i\text{'}\right) = \gamma_{ii}^{t-1}\left(1-\gamma_{ii}\right). \tag{5}$$

The geometric distribution is memoryless, implying that the time until the next transition out of the current state is independent of the time spent in the state.

Langrock and Zucchini (2011) showed how an HMM can be structured to fit any sojourn time distribution with arbitrary precision by mapping multiple latent states to the same output state. The distribution of sojourn times is then a mixture of geometric distributions, which is a phase-type distribution, and the Markov property is transferred to the embedded Markov chain as in an HSMM.[2] Phase-type distributions can be used to approximate any positive-valued distribution with arbitrary precision (Nielsen, 2013). Similarly, time-varying transition probabilities will lead to non-geometrically distributed sojourn times. An estimation approach that allows the transition probabilities to be changing over time, therefore, has the flexibility to fit sojourn time distributions other than the geometric.

Long memory and regime switching

Granger and Teräsvirta (1999) and Gourieroux and Jasiak (2001) showed how simple non-linear time series models with infrequent regime switching can generate a long-memory effect in the autocorrelation function. Around the same time, Diebold and Inoue (2001) showed analytically how stochastic regime switching is easily confused with long memory. They specifically showed that under the assumption that the persistence of the states converges to one as a function of the sample size, the variances of partial sums of a Markov-switching process will match those of a fractionally integrated process. This led Baek et al. (2014) to question the relevance of the HMM for long memory as they found common estimators of the long-memory parameter to be extremely biased when applied to data generated by the Markov-switching model of Diebold and Inoue (2001). Baek et al. (2014) argued that the HMM should be viewed as a short-memory model with some long-memory features rather than a long-memory model.

Gourieroux and Jasiak (2001) emphasized that the distinction between a short-memory model with long-memory features and a long-memory model has important practical implications, for example, when the model is used for making predictions. If a fractional model is retained, the predictions should be based on a long history of the observed series. If, on the other hand, a short-memory (regime-switching) model with long-memory features is selected, then the predictions should be based on only the most recent observations.

Several studies have documented how structural changes in the unconditional variance can cause

[1] The other eigenvalue of $\mathbf{\Gamma}$ is $\lambda = 1$.

[2] In an HSMM, the sojourn time distribution is modeled explicitly for each state. The conditional independence assumption for the observation process is similar to a simple HMM, but the Markov property is transferred to the embedded first-order Markov chain, that is, the sequence of visited states (Bulla and Bulla, 2006).

long-range dependence in the volatility and integrated GARCH effects (see Mikosch and Stărică, 2004, and the references therein). The generalized autoregressive conditional heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986) has been extended in various ways since its introduction in an effort to capture long-range dependencies in economic time series (see e.g. Baillie et al., 1996; Bauwens et al., 2014). Stărică and Granger (2005) identified intervals of homogeneity where the non-stationary behavior of the S&P 500 series can be approximated by a stationary model. They found that the most appropriate is a simple model with no linear dependence but with significant changes in the mean and variance of the time series. On the intervals of homogeneity, the data is approximately a white noise process. Their results indicate the time-varying unconditional variance as the main source of non-stationarity in the S&P 500 series.

Calvet and Fisher (2004) showed that a multi-frequency regime-switching model is able to generate substantial outliers and capture both the low-frequency regime shifts that cause abrupt volatility changes and the smooth autoregressive volatility transitions at mid-range frequencies without including GARCH components or heavy-tailed conditional distributions. The multi-frequency regime-switching model reproduces the long memory of the volatility by having a component with a duration of the same order as the sample size. With two possible values for the volatility the number of states increases at the rate 2^f, where f is the number of switching frequencies. Thus, there are over 1,000 states when $f = 10$.

It was already known that a Markov chain with a countably-infinite state space can have the long-memory property (see Granger and Teräsvirta, 1999, and references therein). The model proposed in this paper is much simpler and, consequently, less likely to be overfitted out of sample or in an on-line application like adaptive forecasting. The model is a simple Gaussian HMM with parameters that are time-varying in a non-parametric way. This approach, in principle, allows for an infinite number of states, but the number of parameters that has to be estimated remains unchanged compared to an HMM with constant parameters. The time variation is observation driven based on the score function of the predictive model density. This is similar to the GAS model of Creal et al. (2013), but the time variation is not limited to the transition probabilities as in the study by Bazzi et al. (2017).

Adaptive parameter estimation

The parameters of an HMM are often estimated using the Maximum Likelihood (ML) method. The likelihood function of an HMM is, in general, a complicated function of the parameters with several local maxima and in mixtures of continuous distributions the likelihood can be unbounded in the vicinity of certain parameter combinations.[3] The two most popular approaches to maximizing the likelihood are direct numerical maximization and the Baum–Welch algorithm, a special case of the Expectation Maximization (EM) algorithm (Baum et al., 1970; Dempster et al., 1977).

When maximizing the likelihood, every observation is typically assumed to be of equal importance no matter how long the sample period is. This approach works well when the sample period is short and the underlying process does not change over time. The time-varying behavior of the parameters documented in previous studies (Rydén et al., 1998; Bulla, 2011), however, calls for an

[3]If, for example, the conditional distribution is Gaussian then the likelihood can be made arbitrarily large by setting the mean equal to one of the observations and letting the conditional variance tend to zero (Frühwirth-Schnatter, 2006).

adaptive approach that assigns more weight to the most recent observations while keeping in mind past patterns at a reduced confidence.

As pointed out by Cappé et al. (2005), it is possible to evaluate derivatives of the likelihood function with respect to the parameters for virtually any model that the EM algorithm can be applied to. This is, for example, true when the standard (Cramer-type) regularity conditions for the ML estimator hold because the maximizing quantities in the M-step are derived based on the derivatives of the likelihood function. As a consequence, instead of resorting to a specific algorithm such as the EM algorithm, the likelihood can be maximized using gradient-based optimization methods. Lystig and Hughes (2002) described an algorithm for exact computation of the score vector and the observed information matrix in HMMs that can be performed in a single pass through the data. The algorithm was derived from the forward–backward algorithm.

The reason for exploring gradient-based methods is the flexibility to make the estimator recursive and adaptive.[4] The estimation of the parameters through a maximization of the conditional log-likelihood function can be done recursively using the estimator

$$\hat{\theta}_t = \arg\max_{\theta} \sum_{n=1}^{t} w_n \log \Pr\left(X_n \left| \boldsymbol{X}^{(n-1)}, \theta \right.\right) = \arg\max_{\theta} \tilde{\ell}_t\left(\theta\right) \tag{6}$$

with $w_n = 1$. The recursive estimator can be made adaptive by introducing a different weighting. A popular choice is to use exponential weights $w_n = \lambda^{t-n}$, where $0 < \lambda < 1$ is the forgetting factor (see e.g. Madsen, 2008). The speed of adaption is then determined by the effective memory length

$$N_{eff} = \frac{1}{1-\lambda}. \tag{7}$$

Maximizing the second order Taylor expansion of $\tilde{\ell}_t\left(\theta\right)$ around $\hat{\theta}_{t-1}$ with respect to θ and defining the solution as the estimator $\hat{\theta}_t$ leads to

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \left[\nabla_{\theta\theta}\tilde{\ell}_t\left(\hat{\theta}_{t-1}\right)\right]^{-1} \nabla_{\theta}\tilde{\ell}_t\left(\hat{\theta}_{t-1}\right). \tag{8}$$

This is equivalent to a specific case of the GAS model of Creal et al. (2013). Using the estimator (8) it is possible to reach quadratic convergence whereas the GAS model in general converges only linearly (see Cappé et al., 2005).

For HMMs, the score function must consider the previous observations and cannot reasonably be approximated by the score function of the latest observation as it is often done for other models (Khreich et al., 2012). In order to compute the weighted score function the algorithm of Lystig and Hughes (2002) has to be run for each iteration and the contribution of each observation has to be weighted. Experimentation suggests that with an effective memory length of 250 observations, it is necessary to compute the contribution of the last 2,500 observations to get a satisfactory approximation of the weighted score function. This leads to a significant increase in computational complexity.

[4] See Khreich et al. (2012) for a survey of techniques for incremental learning of HMM parameters.

Approximating the Hessian by

$$\begin{aligned}\nabla_{\theta\theta}\tilde{\ell}_t\left(\hat{\theta}_{t-1}\right) &= \nabla_{\theta\theta}\sum_{n=1}^{t}\lambda^{t-n}\log\Pr\left(X_n\left|\boldsymbol{X}^{(n-1)},\hat{\theta}_{t-1}\right.\right)\\ &= \sum_{n=1}^{t}\lambda^{t-n}\nabla_{\theta\theta}\log\Pr\left(X_n\left|\boldsymbol{X}^{(n-1)},\hat{\theta}_{t-1}\right.\right)\\ &\approx \sum_{n=1}^{t}\lambda^{t-n}\left(-I_t\left(\hat{\theta}_{t-1}\right)\right)\\ &= \frac{1-\lambda^t}{1-\lambda}\left(-I_t\left(\hat{\theta}_{t-1}\right)\right),\end{aligned} \tag{9}$$

leads to the recursive, adaptive estimator

$$\hat{\theta}_t \approx \hat{\theta}_{t-1} + \frac{A}{\min\left(t, N_{eff}\right)}\left[I_t\left(\hat{\theta}_{t-1}\right)\right]^{-1}\nabla_\theta\tilde{\ell}_t\left(\hat{\theta}_{t-1}\right), \tag{10}$$

where the tuning constant A can be adjusted to increase or decrease the speed of convergence without changing the effective memory length. In order to improve the clarity, the fraction $\frac{1-\lambda}{1-\lambda^t}$ was replaced by $\frac{1}{\min(t,N_{eff})}$, where N_{eff} is the effective memory length (7). The two fractions share the property that they decrease toward a constant when t increases. It is necessary to apply a transformation to all constrained parameters for the estimator (10) to converge. Furthermore, in order to avoid very large initial steps, it is often advisable to start the estimation at a $t_0 > 0$.

The Fisher information can be updated recursively using the identity $\mathrm{E}\left[\nabla_{\theta\theta}\ell_t\right] = -\mathrm{E}\left[\nabla_\theta\ell_t\nabla_\theta\ell_t'\right]$:

$$\begin{aligned}I_t\left(\hat{\theta}\right) &= -\frac{1}{t}\sum_{n=1}^{t}\nabla_\theta\ell_n\left(\hat{\theta}\right)\nabla_\theta\ell_n\left(\hat{\theta}\right)'\\ &= -\frac{t-1}{t}\frac{1}{t-1}\left\{\sum_{n=1}^{t-1}\nabla_\theta\ell_n\left(\hat{\theta}\right)\nabla_\theta\ell_n\left(\hat{\theta}\right)'\right.\\ &\quad\left.+\ \nabla_\theta\log\Pr\left(X_t\left|\boldsymbol{X}^{(t-1)},\hat{\theta}\right.\right)\nabla_\theta\log\Pr\left(X_t\left|\boldsymbol{X}^{(t-1)},\hat{\theta}\right.\right)'\right\}\\ &= I_{t-1}\left(\hat{\theta}\right) + \frac{1}{t}\left[\nabla_\theta\log\Pr\left(X_t\left|\boldsymbol{X}^{(t-1)},\hat{\theta}\right.\right)\right.\\ &\quad\left.\cdot\nabla_\theta\log\Pr\left(X_t\left|\boldsymbol{X}^{(t-1)},\hat{\theta}\right.\right)' - I_{t-1}\left(\hat{\theta}\right)\right].\end{aligned} \tag{11}$$

Further, the matrix inversion lemma is applicable, since the estimator (10) only makes use of the inverse of the Fisher information. The diagonal elements of the inverse of the Fisher information provide uncertainties of the parameter estimates as a by-product of the algorithm.[5]

[5]If some of the parameters are on or near the boundary of their parameter space, which is often the case in HMMs, the use of the Hessian to compute standard errors is unreliable. Moreover, the conditions for asymptotic normality of the ML estimator are often violated, thus making confidence intervals based on the computed standard errors unreliable. In those cases confidence intervals based on the profile likelihood function or bootstrapping provide a better approximation.

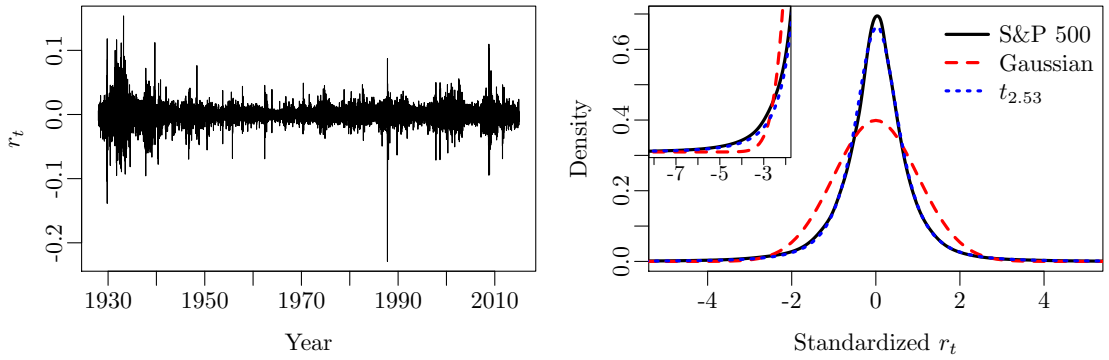

Figure 1. Daily log-returns of the S&P 500 index and the density of the standardized daily log-returns together with the density function for the standard normal distribution and a t-distribution.

Data

The data analyzed is 21,851 daily log-returns of the S&P 500 index covering the period from 1928 to 2014.[6] The log-returns are calculated using $r_t = \log(P_t) - \log(P_{t-1})$, where P_t is the closing price of the index on day t and log is the natural logarithm. It is evident from the plot of the log-returns shown in Figure 1 that the data includes a larger number of exceptional observations. The Global Financial Crisis of 2007–2008 does not stand out when compared to the Great Depression of 1929–1933 and the period around Black Monday in October 1987. The tendency for the volatility to form clusters as large price movements are followed by large price movements and vice versa is a stylized fact (see e.g. Cont, 2001; Lindström et al., 2015).

The excess kurtosis is evident from the plot of the density function shown in Figure 1. There is too much mass centered right around the mean and in the tails compared to the normal distribution. The t-distribution with 2.53 degrees of freedom is a much better fit to the unconditional distribution of the log-returns. There are 169 observations that deviate more than four standard deviations from the mean compared to an expectation of 1.4 observations if the returns were normally distributed. The most extreme being the -22.9% log-return on Black Monday which deviates more than 19 standard deviations from the sample mean.

In Figure 2, the sample autocorrelation function of the squared log-returns is shown together with the ACF of the squared, outlier-corrected log-returns (the two top panels). The dashed lines are the upper boundary of an approximate 95% confidence interval under the null hypothesis of independence (see Madsen, 2008). To analyze the impact of outliers, values outside the interval $\bar{r}_t \pm 4\widehat{\sigma}$, where $\bar{r}_t$ and $\widehat{\sigma}$ are the estimated sample mean and standard deviation, are set equal to the nearest boundary following the approach by Granger and Ding (1995a). According to Granger et al. (2000), the choice of four standard deviations was arbitrary, but experimentation suggested that the results were not substantially altered if the value was changed somewhat. The long memory of the squared returns is evident from both plots although the large number of exceptional observations greatly reduces the magnitude of the ACF of the unadjusted squared returns (see Chan, 1995).

The persistence of the ACF of the squared returns is, to some extent, a consequence of the volatility clustering seen in Figure 1, but the significantly positive level at lags well over 100 is more likely caused by the data-generating process being non-stationary. This is supported by

[6]The definition of the index was changed twice during the period. In 1957, the S&P 90 was expanded to 500 stocks and became the S&P 500 index. The 500 stocks contained exactly 425 industrials, 25 railroads, and 50 utility firms. This requirement was relaxed in 1988.

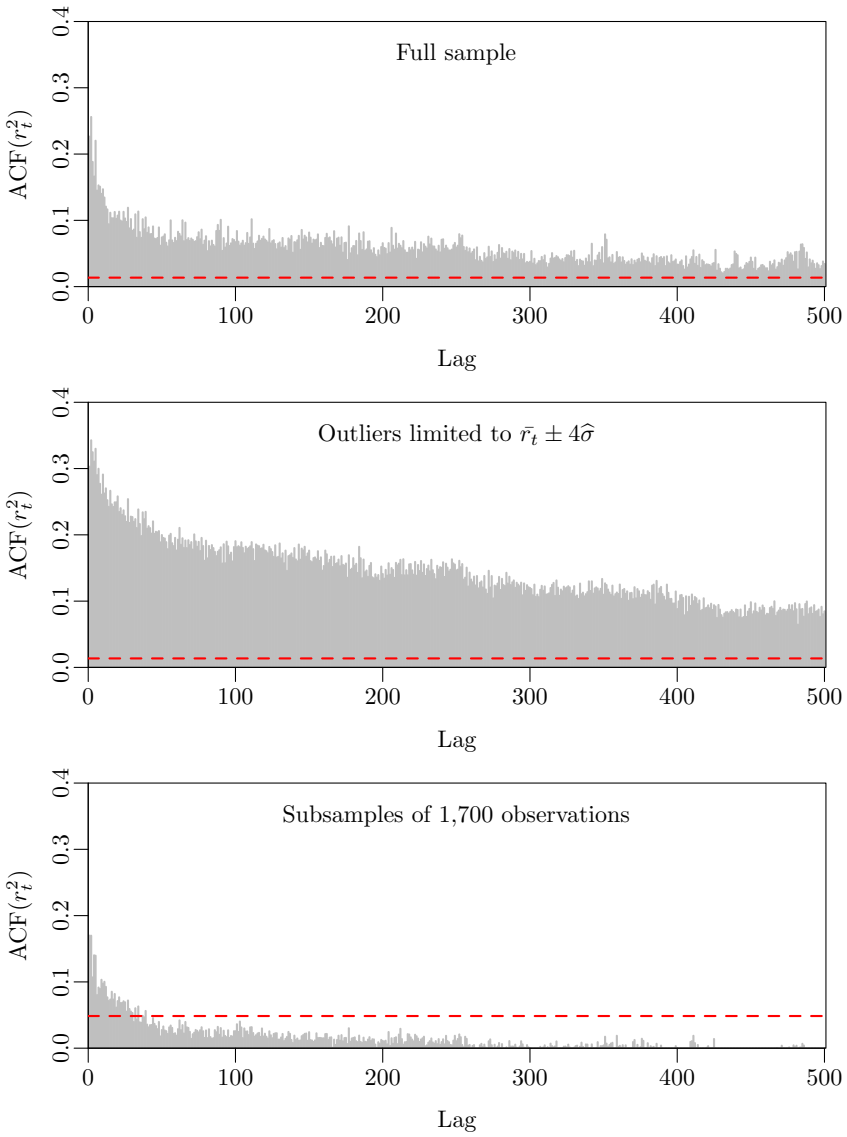

Figure 2. The top panel shows the autocorrelation function of the squared log-returns at lag 1–500. The middle panel shows the autocorrelation function of the squared, outlier-corrected log-returns. The bottom panel shows the average autocorrelation of the squared log-returns for subsamples of 1,700 observations.

the third panel of Figure 2, which shows the average autocorrelation of the squared returns for 12 subsamples of 1,700 observations. The decay of the autocorrelations in the subseries is, on average, substantially faster than in the full series and roughly exponential, as concluded by Malmsten and Teräsvirta (2010). Stărică and Granger (2005) reached the same conclusion based on scaling the absolute log-returns with the time-varying unconditional variance.

Results

Rather than dividing the full sample into subsamples of 1,700 observations, as done by Granger and Ding (1995a), Rydén et al. (1998), and Bulla (2011), the parameters of a two-state HMM with conditional normal distributions are estimated using a rolling window of 1,700 trading days. By using a rolling window it is possible to get an idea of the evolution of the model parameters over time. The result is shown in Figure 3, where the dashed lines are the maximum likelihood estimate (MLE) for the full sample and the gray areas are bootstrapped 95% confidence intervals based on re-estimating the model to 1,000 simulated series of 1,700 observations.

It is evident that the parameters are far from constant as noted by Rydén et al. (1998) and Bulla (2011). The variation of the variances and the transition probabilities far exceeds the likely range of variation as indicated by the 95% confidence intervals. This implies that the sojourn time distribution is not memoryless. The first ten years after World War II, in particular, stand out as the probability of staying in the same state was extraordinarily low. The impact of the extreme returns in 1987 on the parameters of the high-variance state also stands out. One of the drawbacks of using fixed-length forgetting appears from the estimated variance in the second state; the variance spikes on Black Monday in 1987 and does not return to its pre-1987 level until 1,700 days later, thus making the length of the rolling window evident from the figure.

The substantial variation of the parameters could indicate that a regime-switching model is unsuitable for the S&P 500 series, but the model's ability to reproduce the stylized facts suggests otherwise (see Rydén et al., 1998; Bulla, 2011; Nystrup et al., 2015b). It could also be an indication that the model is misspecified, i.e., that it has too few states or a wrong kind of conditional distributions. Rydén et al. (1998) found that in some periods there was a need for a third so-called outlier state with a low unconditional probability. Adding a third state, however, does not lead to smaller variations which suggests that the need for a third state was a consequence of the lack of adaption of the parameters. The addition of a conditional t-distribution in the high-variance state, on the other hand, dampens the variations as shown in Figure 4. The parameters are still not constant, but the variations are smaller especially around 1987. It is only the degrees of freedom of the t-distribution that change dramatically from a maximum of 29 in 1978 to a minimum of 2.5 in 1990.

The choice of window length affects the parameter estimates and can be viewed as a tradeoff between bias and variance. A shorter window yields a faster adaption to changes but a more noisy estimate as fewer observations are used for the estimation. The large fluctuations in the parameter values could be a result of the window length being too short, but the parameter values do not stabilize even if the window length is increased to 2,500 days. Instead, there is a strong incentive to reduce the window length to secure a faster adaption to the non-stationary behavior of the data-generating process. If the window length is reduced to 1,000 days, then the degrees of freedom of the t-distribution exceed 100 throughout a large part of the period, meaning that the distribution in the high-variance state is effectively normal. This suggests that the t-distribution,

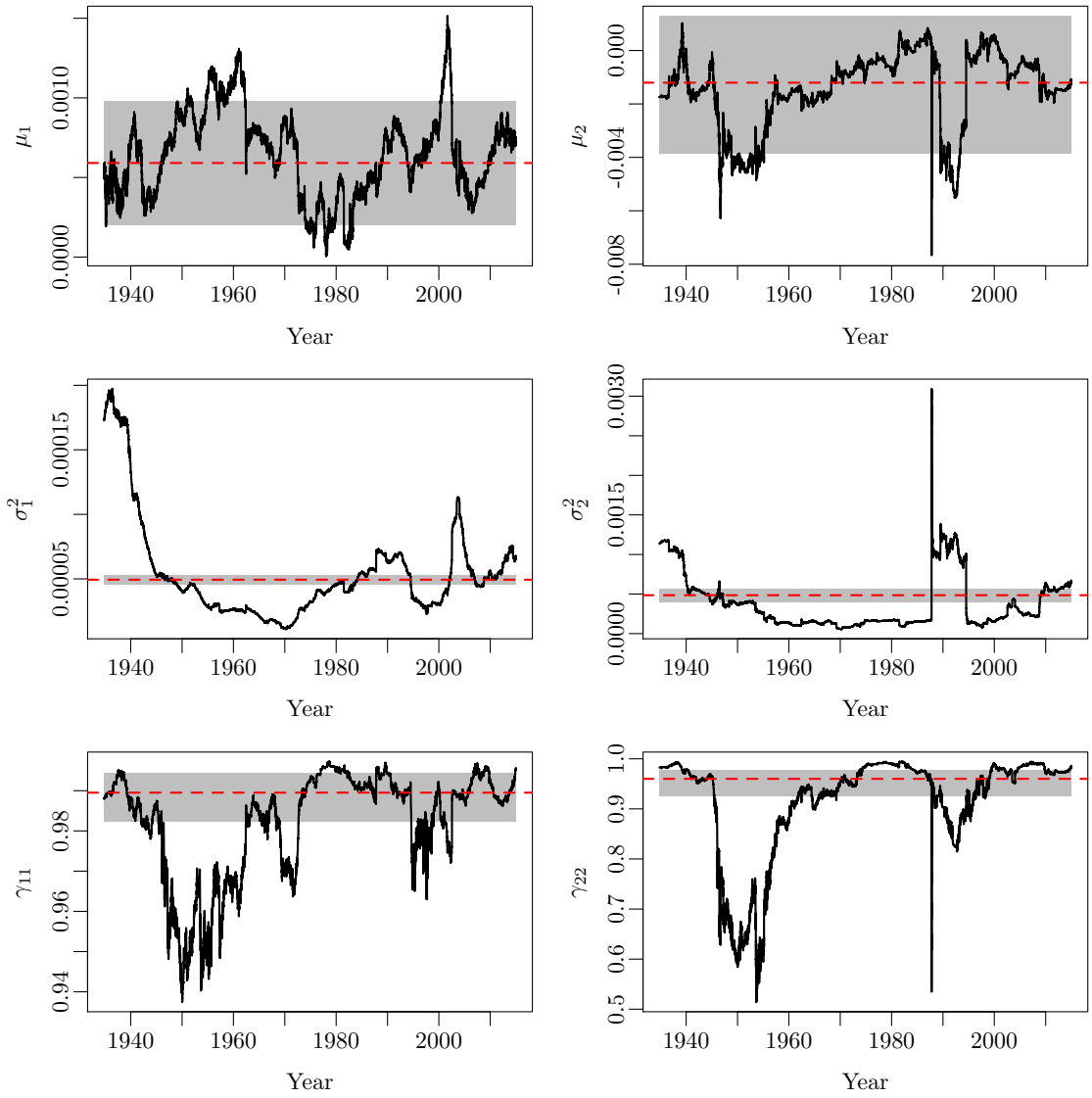

Figure 3. The parameters of a two-state Gaussian HMM estimated using a rolling window of 1,700 trading days. The dashed lines are the MLE for the full series and the gray areas are bootstrapped 95% confidence intervals.

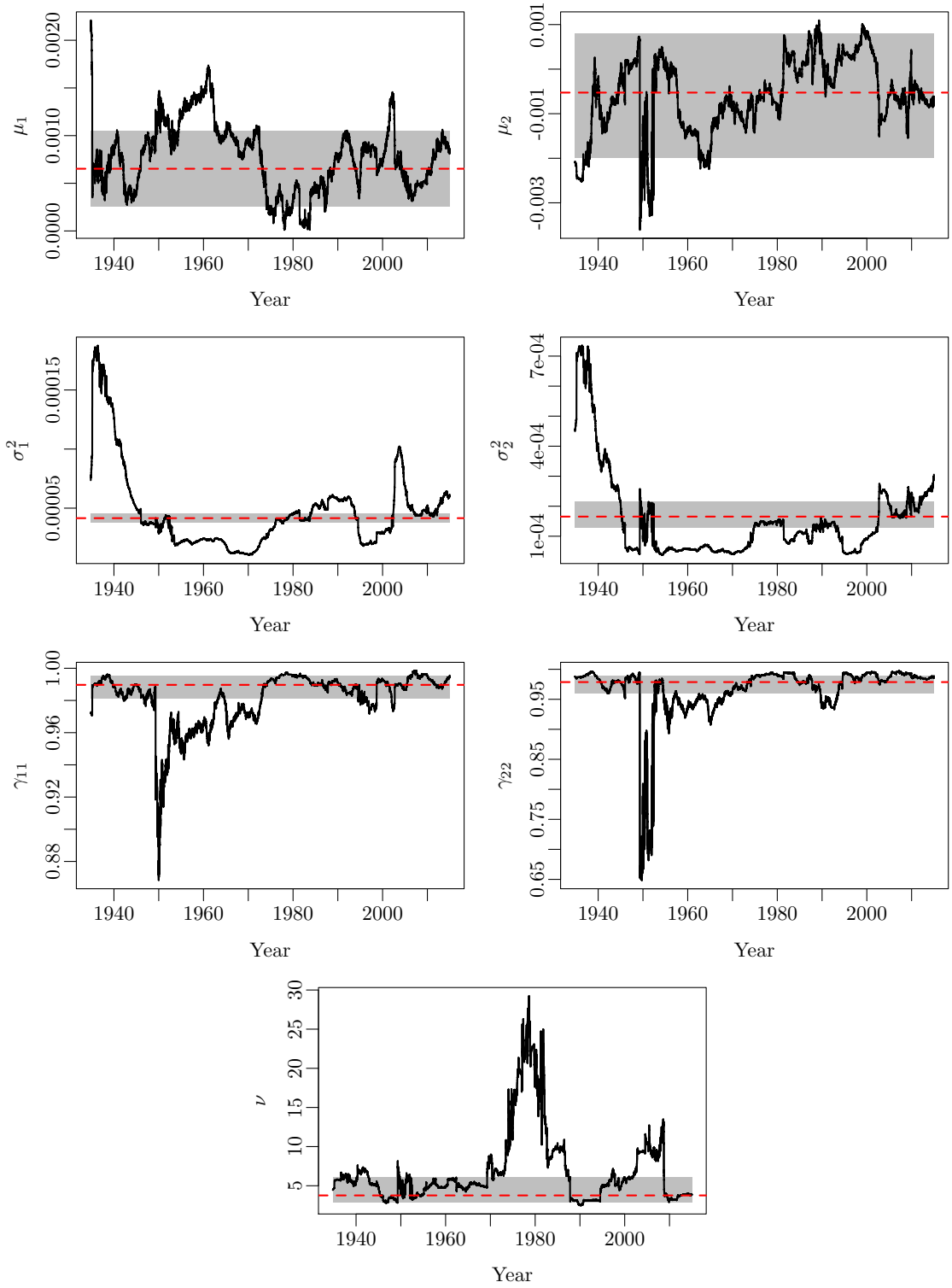

Figure 4. The parameters of a two-state HMM with a conditional t-distribution in the high-variance state estimated using a rolling window of 1,700 trading days. The dashed lines are the MLE for the full series and the gray areas are bootstrapped 95% confidence intervals.

to some extent, is a compensation for too slow an adaption of the parameters.

It cannot be ruled out that adding a couple hundred more states or switching frequencies, as proposed by Calvet and Fisher (2004), would stabilize the parameters, but it would also make the model more likely to be overfitting out of sample and impossible to use for state inference. Given the non-stationary behavior of the data-generating process, it is reasonable to assume that a model with non-constant parameters is needed.

Figure 5 shows the parameters of the two-state HMM with conditional normal distributions estimated using the adaptive estimation approach outlined in Section 4 with an effective memory length $N_{eff} = 250$.[7] The dashed lines show the MLE for the full series and the gray areas are approximate 95% confidence intervals based on the profile likelihood functions. The width of the confidence intervals is seen to spike whenever there are large movements in the parameter estimates.

Compared to Figure 3, the variations are larger as a result of the shorter memory length. Using exponential forgetting, the effective memory length can be reduced compared to when using fixed-length forgetting without increasing the sensitivity to noise to an undesirable level. Exponential forgetting is also more meaningful as it assigns most weight to the newest observations and, at the same time, observations are not excluded from the estimation from one day to the next. This leads to smoother parameter estimates. The optimal memory length depends on the intended application and is not necessarily constant over time, however, 250 days is close to being the minimum length that offers a reasonable tradeoff between speed of adaption and sensitivity to noise.

Reproducing the long memory

The top panel of Figure 6 shows the ACF of the squared log-returns together with the average autocorrelation of squared log-returns based on 100 datasets simulated using the estimated parameters shown in Figure 3–5. The test sample does not include the first 1,700 observations in order to facilitate a comparison with the models estimated using a rolling window. The later starting point causes the ACF to be significantly lower than in Figure 2, supporting the hypothesis that the long memory is caused by non-stationarity.

The autocorrelation of the simulated data is, on average, higher than the sample autocorrelation for all three models. The adaptively estimated Gaussian HMM $\left(\text{HMM}_N^{N_{eff}=250}\right)$ appears to be closest to having the right shape, but it is the HMM with a conditional t-distribution $\left(\text{HMM}_{Nt}^{\text{RW}=1700}\right)$ that provides the best fit to the ACF. The fact that the ACFs of the squared log-returns of the Gaussian HMMs exceed that of the data is an indication that the tails of the Gaussian models are too short compared to the empirical distribution.

When constraining the impact of outliers to $\bar{r}_t \pm 4\widehat{\sigma}$, as done in the second panel of Figure 6, the level of autocorrelation in the simulated data is similar to the empirical data. The adaptively estimated Gaussian HMM provides a good fit to the sample ACF of the squared, outlier-corrected log-returns, while the ACFs of the two models that were estimated using a rolling window are too persistent. The difference in the fit to the ACF of the squared, outlier-corrected log-returns is largest at the highest lags. This observation is supported by the computed mean squared errors for lag 1–250 and 251–500 summarized in Table I. The result when using a threshold of eight instead of four standard deviations is very similar and therefore not shown. With this (equally arbitrary) threshold the model that includes a t-distribution becomes a worse fit at the lowest lags while the

[7] The estimation was started at $t_0 = 250$ in order to avoid very large initial steps with the tuning constant $A = 1.25$. Numerical optimization of the weighted likelihood function was used to find initial values for the parameters and the Hessian.

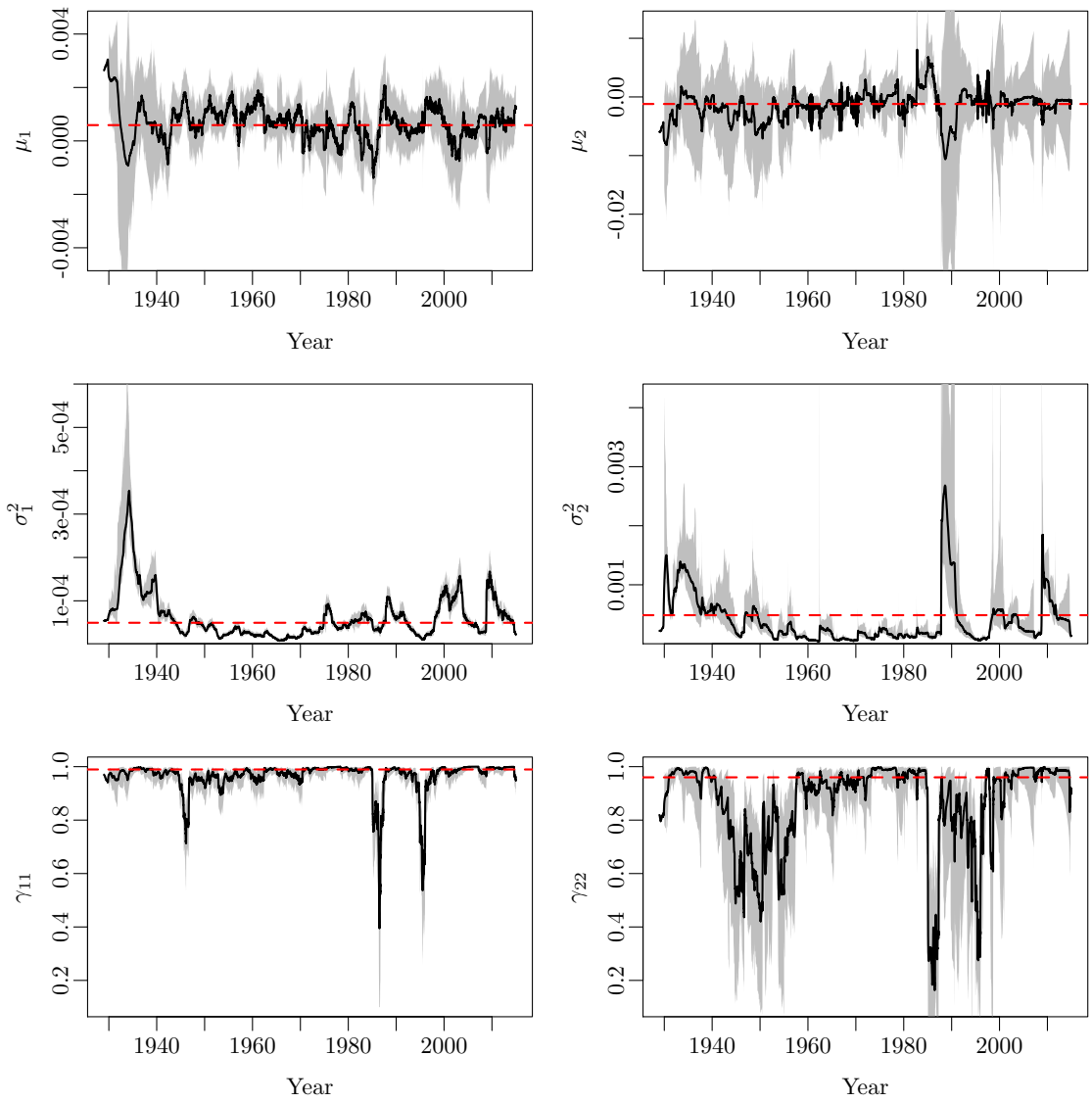

Figure 5. The parameters of a two-state Gaussian HMM estimated adaptively using an effective memory length $N_{eff} = 250$. The dashed lines are the MLE for the full series and the gray areas are approximate 95% confidence intervals based on the profile likelihood functions.

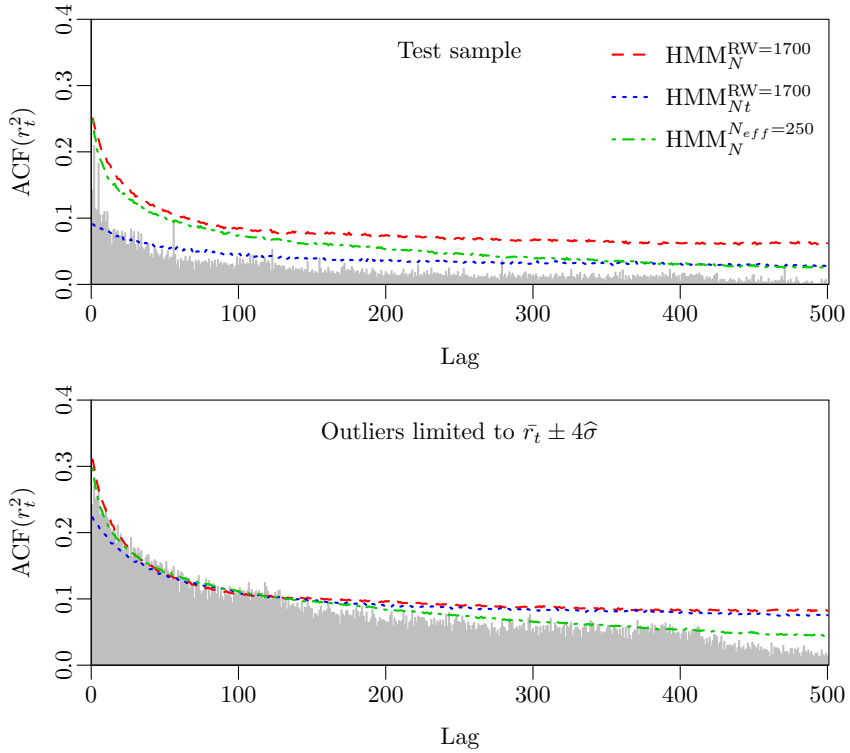

Figure 6. The top panel shows the autocorrelation function of the squared log-returns at lag 1–500 together with the average autocorrelation of squared log-returns simulated using the different estimated models. The test sample does not include the first 1,700 observations. In the bottom panel the impact of outliers is limited to four standard deviations.

Table I. The mean squared error for the ACF of the squared log-returns and the outlier-corrected, squared log-returns at lag 1–250 and 251–500 for the estimated models.

Model	Test sample		Outlier-corrected	
	$\text{MSE}_{1:250} \times 10^3$	$\text{MSE}_{251:500} \times 10^3$	$\text{MSE}_{1:250} \times 10^3$	$\text{MSE}_{251:500} \times 10^3$
$\text{HMM}_N^{\text{RW}=1700}$	3.85	2.99	0.36	1.84
$\text{HMM}_{Nt}^{\text{RW}=1700}$	0.39	0.47	0.32	1.48
$\text{HMM}_N^{N_{eff}=250}$	2.28	0.62	0.17	0.27

adaptively estimated Gaussian HMM still provides the best fit.

Comparing one-step density forecasts

Table II compares the predictive log-likelihood of the different two-state models for the full test sample and when leaving out the 20 most negative contributions to the log-likelihood. This is to separate the impact of the most exceptional observations, similar to the approach used for the ACF. The 20 observations are not the ones furthest away from the sample mean and they are also not the same for all the models, but there is a large overlap. As HMMs are often used for day-to-day state identification in an on-line setting, the focus is on one-step forecasts.

For the full test sample, which does not include the first 1,700 observations, the HMM with a t-distribution in the high-variance state has the highest predictive log-likelihood. When leaving out the 20 observations that make the most negative contributions to the log-likelihood, the adaptively estimated Gaussian HMM outperforms all the other models. This is less than 0.1% of the total

Table II. The predictive log-likelihood of the different two-state models for the full test sample and when leaving out the 20 observations with the most negative contributions.

Model	Predictive log-likelihood	
	Full test sample	Leaving out 20 observations
$\text{HMM}_{N}^{\text{expanding}}$	66,962	67,093
$\text{HMM}_{Nt}^{\text{expanding}}$	67,319	67,386
$\text{HMM}_{N}^{\text{RW=1700}}$	67,391	67,739
$\text{HMM}_{Nt}^{\text{RW=1700}}$	67,757	67,860
$\text{HMM}_{N}^{N_{eff}=250}$	67,678	68,034

number of observations. In fact, the adaptively estimated Gaussian HMM outperforms the other models when removing only the observation from Black Monday. Thus, while the t-distribution is a better fit to the most exceptional observations, the adaptively estimated Gaussian HMM provides the best one-step-ahead density forecasts for the remainder of the sample.

The predictive log-likelihood of both the Gaussian and the combined Gaussian-t model increases when using a rolling rather than expanding window for the estimation. This is not surprising given the non-stationarity of the data-generating process and it clearly shows the need to consider non-stationary models. The difference when leaving out the 20 most negative contributions to the predictive log-likelihood is small when using an expanding window because the tails become very heavy in order to compensate for the lack of adaption of the model parameters. This leads to a poor average predictive performance.

Based on the results shown in Table II, it is natural to wonder whether the performance of the combined Gaussian-t model could be improved by reducing the effective memory length using the adaptive estimation method. This has been attempted, but the advantage of a conditional t-distribution disappears when the memory length is reduced and the degrees of freedom increase. The added uncertainty of the degrees of freedom parameter, which is very sensitive to outliers, leads to more noisy estimates of the mean and scale parameters and a lower predictive likelihood when the memory length is reduced.

To summarize, a shorter memory length of the parameters leads to a good fit to the sample ACF of the squared log-returns when constraining the impact of outliers. A shorter memory also improves the one-step density forecasts with the adaptively estimated Gaussian HMM being the overall best when leaving out the 20 observations that were most difficult to forecast. However, even when the memory length of the parameters is reduced considerably, conditional normal distributions provide a poor fit to those very exceptional observations.

Improving density forecasts with local smoothing

Given that a faster adaption of the parameters leads to improved density forecasts, it should be possible to further improve the density forecasts by improving the parameter forecasts. Recall that the adaptively estimated parameters are found as solutions to (6) for distinct values of t. The observations up to and including time t are used when estimating θ_t, which is then used for making inference about X_{t+1}. In other words, the parameters are assumed to stay constant from time t to time $t+1$.

If the effective memory length is sufficiently short, the approximation of θ_t as a constant vector near t is good. This would imply, however, that a relatively low number of observations is used to estimate θ_t, resulting in a noisy estimate. On the contrary, a large bias may occur if the effective

Table III. The predictive log-likelihood of the estimated models when using cubic smoothing splines to forecast the parameters.

Model	Predictive log-likelihood	
	Full test sample	Leaving out 20 observations
$\text{HMM}_N^{\text{RW}=1700}$	67,408	67,742
$\text{HMM}_{Nt}^{\text{RW}=1700}$	67,771	67,872
$\text{HMM}_N^{N_{eff}=250}$	67,723	68,058

memory is long. Joensen et al. (2000) showed that if the parameter variations are smooth, then locally to t the elements of θ_t are better approximated by local polynomials than local constants.

The parameter variations displayed in Figure 5 appear to be smooth, especially the variations in the conditional variances. The idea of Joensen et al. (2000) is implemented with more flexible cubic smoothing splines rather than polynomials. Table III summarizes the predictive log-likelihood of the two-state models when using cubic smoothing splines to forecast the parameters.

By fitting a cubic smoothing spline to the last nine parameter estimates and then using the fitted spline to forecast the parameters at the next time step, it is possible to increase the predictive log-likelihood of the adaptively estimated Gaussian HMM from 67,678 to 67,723 for the full test sample. When leaving out the 20 most negative contributions, the predictive log-likelihood is improved from 68,034 to 68,058. Thus, the largest improvement is obtained for the 20 observations that are most difficult to forecast. For the two models that were estimated using a rolling window the improvement in forecasting performance is smaller. This is not surprising given that the parameter variations appeared less smooth when using a rolling window for the estimation.

Modeling the driving forces behind the variations is beyond the scope of this paper, but the improvement in forecasting performance that can be obtained by using simple smoothing splines to forecast the parameters suggests there is great potential for a hierarchical model.

Conclusion

By applying an adaptive estimation method that allowed for observation-driven time variation of the parameters it was possible to reproduce the long memory that is characteristic for long series of squared daily returns with a two-state Gaussian hidden Markov model. The transition probabilities were found to be time varying, implying that the sojourn time distribution was not the memoryless geometric distribution.

The adaptive estimation approach meant that the effective memory length could be reduced compared to when using fixed-length forgetting, thereby allowing a faster adaption to changes and a better reproduction of the current parameter values. This led to an improved fit to the autocorrelation function of the squared log-returns and better one-step density forecasts. A third state or a conditional t-distribution in the high-variance state may be necessary to capture the full extent of excess kurtosis in a few periods, but the long memory that is needed to justify a third state or a conditional t-distribution with long tails is not consistent with the fast adaption of the parameters that led to the best fit to the long memory of the squared log-returns and the best one-step density forecasts with the exception of the most extreme observations.

The presented results emphasize the importance of taking into account the non-stationary behavior of the data-generating process. The longer the data period, the more important this is. The outlined adaptive estimation method can be applied to other models than regime-switching

models and for other purposes than density forecasting. Within financial risk management, for example, possible applications include value-at-risk estimation and dynamic asset allocation based on out-of-sample state identification.

A better description of the time-varying behavior of the parameters is an open route for future work that can be pursued in various ways. One way would be to allow different forgetting factors for each parameter or consider more advanced state-dependent, time-dependent, or data-dependent forgetting. Another way would be to formulate a model for the parameter changes in the form of a hierarchical model possibly including relevant exogenous variables. The proposed method for estimating the time variation of the parameters is an important step toward the identification of a hierarchical model structure.

Acknowledgments

This work was supported by Innovation Fund Denmark under Grant No. 4135-00077B.

References

Baek C, Fortuna N, Pipiras V. 2014. Can Markov switching model generate long memory? *Economics Letters* **124**: 117–121.

Baillie RT, Bollerslev T, Mikkelsen HO. 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **74**: 3–30.

Baum LE, Petrie T, Soules G, Weiss N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41**: 164–171.

Bauwens L, Dufays A, Rombouts JV. 2014. Marginal likelihood for Markov-switching and change-point GARCH models. *J. Econometrics* **178**: 508–522.

Bazzi M, Blasques F, Koopman SJ, Lucas A. 2017. Time varying transition probabilities for Markov regime switching models. *J. Time Series Analysis* .

Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**: 307–327.

Bulla J. 2011. Hidden Markov models with t components. Increased persistence and other aspects. *Quantitative Finance* **11**: 459–475.

Bulla J, Bulla I. 2006. Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics and Data Analysis* **51**: 2192–2209.

Bulla J, Mergner S, Bulla I, Sesboüé A, Chesneau C. 2011. Markov-switching asset allocation: Do profitable strategies exist? *J. Asset Management* **12**: 310–321.

Calvet LE, Fisher AJ. 2004. Regime-switching and the estimation of multifractal processes. *J. Financial Econometrics* **2**: 49–83.

Cappé O, Moulines E, Rydén T. 2005. *Inference in hidden Markov models*. Springer.

Chan WS. 1995. Understanding the effect of time series outliers on sample autocorrelations. *Test* **4**: 179–186.

Cont R. 2001. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance* **1**: 223–236.

Creal D, Koopman SJ, Lucas A. 2013. Generalized autoregressive score models with applications. *J. Applied Econometrics* **28**: 777–795.

Dacco R, Satchell S. 1999. Why do regime-switching models forecast so badly? *J. Forecasting* **18**: 1–16.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society* **39**: 1–38.

Diebold FX, Inoue A. 2001. Long memory and regime switching. *J. Econometrics* **105**: 131–159.

Engle RF. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**: 987–1007.

Frühwirth-Schnatter S. 2006. *Finite Mixture and Markov Switching Models*. Springer.

Gourieroux C, Jasiak J. 2001. Memory and infrequent breaks. *Economics Letters* **70**: 29–41.

Granger CWJ, Ding Z. 1995a. Some properties of absolute return: An alternative measure of risk. *Annales D'Economie Et Statistique* **40**: 67–92.

Granger CWJ, Ding Z. 1995b. Stylized facts on the temporal and distributional properties of daily data from speculative markets. Unpublished paper, Department of Economics, University of California, San Diego.

Granger CWJ, Spear S, Ding Z. 2000. Stylized facts on the temporal and distributional properties of absolute returns: An update. In *Proceedings of the Hong Kong International Workshop on Statistics in Finance*. Imperial College Press, 97–120.

Granger CWJ, Teräsvirta T. 1999. A simple nonlinear time series model with misleading linear properties. *Economics Letters* **62**: 161–165.

Joensen A, Madsen H, Nielsen HA, Nielsen TS. 2000. Tracking time-varying parameters with local regression. *Automatica* **36**: 1199–1204.

Khreich W, Granger E, Miri A, Sabourin R. 2012. A survey of techniques for incremental learning of HMM parameters. *Information Sciences* **197**: 105–130.

Langrock R, Zucchini W. 2011. Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics and Data Analysis* **55**: 715–724.

Lindström E, Madsen H, Nielsen JN. 2015. *Statistics for Finance*. Chapman & Hall.

Lystig TC, Hughes JP. 2002. Exact computation of the observed information matrix for hidden Markov models. *J. Computational and Graphical Statistics* **11**: 678–689.

Madsen H. 2008. *Time Series Analysis*. Chapman & Hall.

Malmsten H, Teräsvirta T. 2010. Stylized facts of financial time series and three popular models of volatility. *European Journal of Pure and Applied Mathematics* **3**: 443–477.

Mikosch T, Stărică C. 2004. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics* **86**: 378–390.

Nielsen BF. 2013. *Matrix Analytic Methods in Applied Probability with a View towards Engineering Applications*. Doctoral thesis, Technical University of Denmark.

Nystrup P, Hansen BW, Madsen H, Lindström E. 2015a. Regime-based versus static asset allocation: Letting the data speak. *The Journal of Portfolio Management* **42**: 103–109.

Nystrup P, Madsen H, Lindström E. 2015b. Stylised facts of financial time series and hidden Markov models in continuous time. *Quantitative Finance* **15**: 1531–1541.

Rydén T, Teräsvirta T, Åsbrink S. 1998. Stylized facts of daily return series and the hidden Markov model. *J. Applied Econometrics* **13**: 217–244.

Stărică C, Granger CWJ. 2005. Nonstationarities in stock returns. *Review of Economics and Statistics* **87**: 503–522.

Zucchini W, MacDonald IL. 2009. *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman & Hall, 2nd edition.